REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-04-

0224

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
07-04-2004	Final	May 1, 2001/Sept 30, 2003

4. TITLE AND SUBTITLE
~~CIP Fellowship in Cyberforensics~~

5a. CONTRACT NUMBER

5b. GRANT NUMBER
F49620-01-1-0264

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Peter P. Chen

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Louisiana State University
Computer Science Department
298 Coates Hall
Baton Rouge, LA 70803

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/PIE
4015 Wilson Blvd. Room 713
Arlington, VA 22203-1954

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

20040426 068

13. SUPPLEMENTARY NOTES

14. ABSTRACT

LSU was one of the universities chosen to participate in the project of training new researchers to work on the Critical Infrastructure Protection and Information Assurance (CIPIA) areas. Three Ph.D.'s (Steve Seiden, Guoli Ding, and Nigel Gwee) who were not in the Cyber Security area were selected to become the CIPIA Fellows at LSU. These fellows were trained intensively using multiple methods: auditing relevant courses, reading relevant books and articles, communicating with CIPIA experts inside and outside of LSU, and participating in CIPIA research projects. Although they have participated in several CIPIA projects, the main thrust has been developing mathematical models and efficient algorithms to identify malicious cyber transactions and terrorists. Each of the three Fellows has produced excellent research results. Steve Seiden has developed several online algorithms useful in cyber security. Guoli Ding and Steve Seiden have developed an anonymous communication scheme based on the concept of randomized busing to protect the security of communication no matter in cyber space or in other environments. Nigel Gwee has developed a technique to combine several heuristics algorithms together to locate the malicious cyber transactions quicker. In summary, this CIPIA Fellow project at LSU has been very successful as indicated by the following facts: (1) a huge number of research papers has been produced by these 3 fellows (Seiden has 16, Ding has 9, and Gwee has 3), and (2) teaming up with his mentor, Ding has obtained a large NSF grant on cyber security as a Co-PI.

15. SUBJECT TERMS
Critical infrastructure protection, information assurance, cyber security, malicious cyber transactions, training.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Peter P. Chen
U	U	U	Unlimited	20	19b. TELEPHONE NUMBER (include area code) (225) 578-2483

Critical Infrastructure Protection and Information Assurance (CIPIA) Fellow Program Final Report

Contract Number: AFOSR Grant No. F49620-01-1-0264
Contract Period: 5/1/01 – 9/30/03

Submitted by

Peter P. Chen, Ph.D.
Foster Distinguished Chair Professor
Computer Science Department
Louisiana State University
Baton Rouge, LA 70803
Tel: (225) 578-2483, Fax: (225) 578-1965
E-mail: pchen@lsu.edu
Web: http://www.csc.lsu.edu/~chen

September 2003

Abstract

LSU was one of the universities chosen to participate in the project of training new researchers to work on the Critical Infrastructure Protection and Information Assurance (CIPIA) areas. Three Ph.D.'s (Steve Seiden, Guoli Ding, and Nigel Gwee) who were not in the Cyber Security area were selected to become the CIPIA Fellows at LSU. These fellows were trained intensively using multiple methods: auditing relevant courses, reading relevant books and articles, communicating with CIPIA experts inside and outside of LSU, and participating in CIPIA research projects. Although they have participated in several CIPIA projects, the main thrust has been developing mathematical models and efficient algorithms to identify malicious cyber transactions and terrorists. Each of the three Fellows has produced excellent research results. Steve Seiden has developed several online algorithms useful in cyber security. Guoli Ding and Steve Seiden have developed an anonymous communication scheme based on the concept of randomized busing to protect the security of communication no matter in cyber space or in other environments. Nigel Gwee has developed a technique to combine several heuristics algorithms together to locate the malicious cyber transactions quicker. In summary, this CIPIA Fellow project at LSU has been very successful as indicated by the following facts: (1) a huge number of research papers has been produced by these 3 fellows (Seiden has 16, Ding has 9, and Gwee has 3), and (2) teaming up with his mentor, Ding has obtained a large NSF grant on cyber security as a Co-PI.

1. Introduction

This is the final report of the Critical Infrastructure Protection and Information Assurance (CIPIA) Fellow Program project at Louisiana State University (LSU). The Principal Investigator (the mentor of the CIPIA Fellows) is Dr. Peter P. Chen, Murphy J. Foster Chair Professor of Computer Science.

Three fellows were selected and trained in this project at LSU. The results of the training are reported here.

In the following, we will first discuss how these three fellows were trained. Then, we will give a detailed report on each of the three fellows. The final section is the conclusion.

2. How the Fellows were Trained

The fellows are given introductory material (books, articles, etc.) to read and discussed their understandings and findings with the mentor. The mentor then guided them to start to do research in topics related to CIPIA. First, started with simple topics and then got into more and more sophisticated topics. We also invited internal and external CIPIA experts to give seminars and to interact with the fellows. The fellows also interacted with the faculty and students who either have done research or are interested in CIPIA topics. We also asked the fellows to give talks on their own research work on CIPIA topics.

3. Detailed Report on Each of the Three Fellows

For each individual Fellow, we provide the same type of detailed information.

3.1. Detailed report on Follow #1: Dr. Steven Seiden

1) PI Name and Contact information:
Dr. Peter P. Chen, Foster Distinguished Chair Professor, Computer Science Dept., Louisiana State University, Baton Rouge, LA 70803; E-Mail: pchen@lsu.edu.

2) Name of Fellow:
Dr. Steve Seiden

3) Contact information:
Dr. Seiden passed away in a tragic accident in June 2002.

4) Background information:
Dr. Seiden was an Assistant Professor in the Computer Science dept of LSU, specializing in Theory of Computation.

5) preFellowship; (Background information on Fellow including degree, field or areas of training, etc.):
Ph.D. in Information and Computer Science, University of California, Irvine, 1996; Thesis title: Randomization in Online Computation. His areas of training and specialties were: Operations Research, Online Algorithms, Randomization, and Scheduling.

6) Fellowship Period; (Brief summary of activities, areas of study and accomplishments during the fellowship period including publications and pending publications):
Dr. Seiden's Fellowship period started in January 2002 (however, his research work on cyber security started from September 2001 with financial support from LSU). Before the fellowship, Dr. Seiden was a theoretical computer scientist. The fellowship has exposed him to many interesting and challenging problems in cyber security that could be solved by his mathematical skills in scheduling, randomization, and online algorithms. In particularly, he concentrated on several areas of research during his fellowship period: (a) fundamental mathematical techniques useful to cyber security, (2) merging of security privilege structures of different computer systems (for example, paper #C1, #C2, and #C8), (3) anonymous communication techniques (Paper #C8), and (4) machine learning techniques for online algorithms (Paper #D2).

The following is a list of papers published and pending:
(A) Papers Published

1. Seiden, S., "On the online bin packing problem," Proceedings of the 28th International Colloquium on Automata, Languages and Programming (ICALP`01) (July 2001), pp. 237-249.

2. Seiden, S., and van Stee, R. "New bounds for multi-dimensional packing," Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA`02) (January 2002), pp. 486-495.
3. Augustine, J., and Seiden, S., "Linear time approximation schemes for vehicle scheduling," Proceedings of the 8th Scandinavian Workshop on Algorithm Theory, (SWAT`02) (July 2002).
4. Epstein, L., Seiden, S., and van Stee, R., "New bounds for variable-sized and resource augmented online bin packing," Proceedings of the 29th International Colloquium on Automata, Languages and Programming (ICALP`02) (July 2002).
5. Seiden, S., "On the online bin packing problem, JACM, Vol. 49, No. 5 (September 2002), pp. 640-671.

(B) Papers Accepted for Publication

1. Seiden, S. "A general decomposition theorem for the k-server problem," Information and Computation.

(C) Papers Submitted for Publication

1. Chen, P.P., Ding, G., and Seiden, S., "Poset Merging with Applications to Database Security."
2. Seiden, S. and Chen, P.P., "New Bounds for Randomized Busing."
3. N. Alon, G. Ding, B. Oporowski, and D. Vertigan, "Partitioning into graphs with only small components," Submitted to: Journal of Combinatorial Theory, Series B.
4. M. DeVos, G. Ding, B. Oporowski, B. Reed, D. Sanders, P. Seymour, and D. Vertigan,"Excluding any graph as a minor allows a low tree-width 2-coloring," Submitted to: Journal of Combinatorial Theory, Series B.
5. Feder, T., Motwani, R., Panigrahy, R., Seiden, S., van Stee, R., and Zhu, A., "Web caching with request reordering," submitted for publication.
6. Fiat, A., Mendel, M., and Seiden, S., "Online companion caching," submitted for publication.
7. Seiden, S., "An improved lower bound for uniform metrical task systems," submitted for publication.
8. Seiden, S., Chen, P.P., Lax, R.F., Chen, J. and Ding, G., "New Bounds for Randomized Busing," submitted for publication.

(D) Working Papers

1. Arkin, E., Bender, M., Bunde, D., Lal, A., Leung, V. J., Mitchell, J. S. B., Phillips, C. A., and Seiden, S. "Methods for improving CPlant through processor allocation."
2. Seiden, S., "How to better use Expert Advice."

7) Current Status and Future plans (tentative or otherwise) of the Fellow (post-Fellowship):

It is sad that Dr. Seiden was killed by a truck when he was riding a bicycle in June 2002. Otherwise, he would have become a productive researcher in CIPIA. What a tragic loss!

3.2. Detailed report on Fellow#2: Dr. Guoli Ding

1) PI Name and Contact information:
Dr. Peter P. Chen, Foster Distinguished Chair Professor, Computer Science Dept., Louisiana State University, Baton Rouge, LA 70803; E-Mail: pchen@lsu.edu.

2) Name of Fellow:
Dr. Guoli Ding

3) Contact information:
Mathematics Dept., Louisiana State University, Baton Rouge, LA 70803.

4) Background information:
Dr. Ding was a mathematician and a professor in the Math dept of LSU.

5) preFellowship; (Background information on Fellow including degree, field or areas of training, etc.):
Ph.D. in Operations Research, Rutgers University, 1991; His areas of training and specialties were: Graph Theory, Combinatorial Optimization, and Operations Research

6) Fellowship Period; (Brief summary of activities, areas of study and accomplishments during the fellowship period including publications and pending publications):
Dr. Ding's Fellowship period started in the summer of 2001. Before the fellowship, Dr. Ding was a pure mathematician. The fellowship has exposed him to many interesting and challenging problems in cyber security that could be solved by his mathematical skills in graph theory and combinatorial optimization. He has become very interested in cyber security problems. He has concentrated on several research directions: (a) mathematical techniques useful in cyber security, and (b) the "profiling problem," that is, to find an efficient technique to identity terrorists from a group of people (or to identify the cyberspace intruders from a large amount of Internet transactions/messages). The following is a list of papers published and pending:

1. Papers Published

- Ding, G., and Chen, P.P., "Generating r-regular Graphs," Discrete Applied Mathematics, Vol. 129, 2003, pp. 329-343.
- Ding, G., "Excluding any graph as a minor allows a low tree-width 2-coloring" (joint with Matt DeVos, Bogdan Oporowski, Bruce Reed, Daniel. Sanders, Paul Seymour, and Dirk Vertigan), Journal of Combinatorial Theory, Series B, 2003.

(B) Papers Accepted for Publications
- Chen, P.P. and Ding, G., "The Best Expert vs. the Smartest Algorithm," to appear in Theoretical Computer Science.
- Ding, G., and Chen, P.P., "Unavoidable Double-Connected Large Graphs," to appear in Discrete Mathematics.

(C) Papers Submitted for Publication

- Chen, P.P., Ding, G., and Seiden, S., "Poset Merging with Applications to Database Security."
- Chen, P. P. and Ding, G., "A Greedy Heuristic for a Generalized Set Covering Problem."
- Ding, G., and Kanno, J., "Splitter theorems for cubic graphs," Submitted to Combinatorics, Probability and Computing.
- Partitioning graphs into two graphs with only small components, (Joint with D. Sanders, B. Oporowski, and D. Vertigan), Submitted to Combinatorica.
- Seiden, S., Chen, P.P., Lax, R.F., Chen, J. and Ding, G., "New Bounds for Randomized Busing," submitted for publication.

7) Current Status and Future plans (tentative or otherwise) of the Fellow (post-Fellowship):

Dr. Ding is a shinning example of the success of the CIPIA Fellow program. From a pure mathematician, he now is a Co-PI of a large NSF research grant ($1.8 million dollars) on "Cyber security and anti-terrorism" (NSF Grant number #0326387. Also, he was promoted from Associate Professor of Mathematics to Professor of Mathematics.

Dr. Ding intends to continue to pursue research activities in CIPIA using his mathematical skills at his current position as Professor of Mathematics at LSU. He is currently writing papers and proposals with his mentor, Dr. Chen, to continue the research in cyber security, particularly the profiling problem.

3.3. Detailed report on Fellow#3: Dr. Nigel Gwee

1) PI Name and Contact information:
Dr. Peter P. Chen, Foster Distinguished Chair Professor, Computer Science Dept., Louisiana State University, Baton Rouge, LA 70803; E-Mail: pchen@lsu.edu.

2) Name of Fellow:
Dr. Nigel Gwee

3) Contact information:
Computer Science Dept., Louisiana State University, Baton Rouge, LA 70803.

4) Background information:
Dr. Gwee was a scholar in Musicology and an instructor in the computer science department at LSU.

5) preFellowship: (Background information on Fellow incluGwee degree, field or areas of training, etc.):
Ph.D. in Musicology, LSU, 1996. Originally, he planned to become a CIPIA Fellow in 2001, but his citizen naturalization process was delayed by the 9-11 terrorist attacks. Subsequently, he received a second Ph.D. degree in 2002 from LSU in computer science. He became a U.S. citizen in late 2002. His areas of training and specialties were: Computational Complexity in Musicology.

6) Fellowship Period; (Brief summary of activities, areas of study and accomplishments during the fellowship period including publications and pending publications):

Dr. Gwee's Fellowship period started in February 2003, but he actually started to work informally with his mentor, Dr. Peter Chen, since September 2002. Before the fellowship, Dr. Gwee was a musician and then a mathematician with applications to musicology. The fellowship has exposed him to many interesting and challenging problems in cyber security. Currently, he is working on the comparisons of different algorithms for the "profiling problem," that is, to compare different technique to identity terrorists from a group of people (or to identify the cyberspace intruders from a large amount of Internet transactions/messages). The following is a list of papers published and pending:

(A) Papers Published

- Gwee, N., "Composing species counterpoint with genetic algorithms," Proceedings of 41st ACM Southeast Regional Conference, Savannah, GA, March 7-8, 2003, pp. 235-240.

(B) Working Papers

- Gwee, N., and Chen, P.P., "Comparisons of Several Algorithms for a Generalized Set Covering Problem using Simulations."

- Gwee, N., and Chen, P.P., "The whole greater than the sum of its parts: combining the strengths of heuristic optimization algorithms,"

7) Current Status and Future plans (tentative or otherwise) of the Fellow (post-Fellowship):
Dr. Ding is another shinning example of the success of the CIPIA Fellow program. From a music Ph.D., he now is part-time post-doctor of a large NSF research grant ($1.8 million dollars) on "Cyber security and anti-terrorism" (NSF Grant number #0326387. Also, he is an instructor in the Computer Science Department of LSU incorporating some of the CIPIA topics and research results into his course contents.

Dr. Gwee intends to continue to pursue research activities in CIPIA using his skills developed during this Fellowship. He intends to continue to be a postdoctoral researcher working with his mentor, Dr. Peter Chen, to perform research in cyber security problems.

4. Selected Research Papers Produced by the CIPIA Fellow

In this section, we attach the first few pages of several selected papers of the CIPIA Fellows of this project to give you an idea of the quality and type of research work they have done during the project period:

- **Seiden, S.**, Chen, P.P., Lax, R.F., Chen, J. and **Ding, G.**, "New Bounds for Randomized Busing," submitted for publication. Dr. Seiden and Dr. Ding are two of the CIPIA Fellows of this project. This paper addresses a very critical problem in communication, that is, how to hide both the content and the fact of communication from your adversaries. The problem happens in many different forms in military and civilian environment. For example, if we want to communicate with a secret agent in a hostile country, not only need we to hide the content of our message to the agent but also the fact that we are communicating with the agent. Another example is that the mobile missile locations need to be changed constantly. To avoid the surveillance from the satellites, we need to disguise the movement of the missiles so that our adversaries do not know whether we are moving the real missiles or not. In this paper, the CIPIA Fellows Seiden and Ding, together with other researchers, proposed a novel approach based on the "randomized busing" scheme so that the adversaries have no way to know the intended destinations of the buses and also the contents of the buses.
- Chen, P.P. and **Ding, G.**, "The Best Expert vs. the Smartest Algorithm," to appear in Theoretical Computer Science. Dr. Ding is a CIPIA Fellow of this project. This paper addresses a critical problem in machine learning, which has significant applications in cyber security. For example, when we have collected a large amount of cyber traffic data, how can we narrow down the malicious transactions? In machine learning community, there have always been debates on whether human experts can perform better than the best algorithms? The CIPIA Fellow Ding and his mentor Chen studied this problem and identified the performance bounds/differences of these two approaches under certain environments.
- **Gwee, N.**, and Chen, P.P., "The whole greater than the sum of its parts: combining the strengths of heuristic optimization algorithms," Dr. Gwee is a CIPIA Fellow. This problem was triggered by our study of the "profiling techniques" to identify malicious cyber transactions and terrorists. The problem was modeled as a set covering problem, and we have considered different heuristics techniques to find the optimal solution. During this study, CIPIA Fellow Gwee and his mentor Chen discovered new ways to combine several heuristics techniques to improve the speed of finding the best solution.

For each of these three research papers, we include the first 3 pages of the papers in this section. For those who are interested in getting the full-length papers, please contact the P.I. of this project or the CIPIA Fellows directly.

New Bounds for Randomized Busing *

Steven S. Seiden[a] Peter P. Chen[a†‡] R. F. Lax[b] J. Chen[a] Guoli Ding[b]

[a] Department of Computer Science, 298 Coates Hall, LSU, Baton Rouge, LA 70803
[b] Department of Mathematics, LSU, Baton Rouge, LA 70803

Abstract

We consider anonymous secure communication, where parties not only wish to conceal their communications from outside observers, but also wish to conceal the very fact that they are communicating. We consider the bus framework introduced by Beimel and Dolev [2], where messages are delivered by a bus traveling on a random walk. We generalize this idea to consider more than one bus. We show that if w buses are allowed, then the expected delivery time for a message can be decreased from $\Theta(n)$ to $\Theta(n/\sqrt{w})$ in the case of a complete graph. Additionally, we introduce a class of graphs called r-partite directed collars and obtain analogous bounds on the expected delivery time for these graphs. We also propose several new features that resolve possible shortcomings in the systems proposed by Beimel and Dolev.

1 Introduction

Suppose we have a communication network, modeled by a graph G, composed of n vertices and m edges (or arcs, in the case of a directed graph). Messages are passed through this network, so that the various nodes can communicate with each other. A well-studied problem is that of how to encrypt messages, so that even if an outside observer is able to intercept messages, the information being passed remains secret. A different and less well-studied problem is the following: Suppose we wish to conceal not only the contents of a message, but its point of origin and destination. We might imagine that the communications network is a military network for country A, over which critical orders are transmitted. We might wish to conceal which node is the command center, so that an enemy, say country B, does not know where to attack. Further, we may wish to conceal the fact that orders of some kind are being transmitted, as this may alert country B to a coming attack from A. This is known as the *anonymous communication* problem.

Previous Results: The anonymous communication problem was first explored by Chaum, who proposed and analyzed a basic approach called a *mix* [5]. Mixes are further explored in [15, 16, 17]. Another approach to anonymous communication is to use generic secure multi-party function evaluation [3, 4, 7, 6, 12]. However, such schemes can be very inefficient [2]. To solve some of the problems with these methods, two further schemes have been proposed. The first is the *xor-tree* scheme developed by Dolev and Ostrovsky [10]. The second is the *bus* scheme introduced by Beimel and Dolev [2]. In this paper, we focus on the bus scheme.

*Research supported by AFOSR grant No. F49620-01-1-0264 and NSF grant No. 0326387.
†Corresponding author
‡*Email address:* chen@bit.csc.lsu.edu

Beimel and Dolev actually propose several different busing schemes. These schemes can be classified as either *deterministic* or *randomized.* Their main focus is on deterministic schemes, whereas our main focus shall be on randomized schemes. A drawback of the deterministic schemes of Beimel and Dolev is as follows: In all of the deterministic protocols proposed by these authors, the route a message takes through the network is fixed. If an enemy cuts a particular edge, or corrupts messages at a particular node, this could lead to the situation where the communication path between two nodes is unusable. The protocols have no possibility of exploring alternative paths. Essentially, in these protocols, it is possible to discern the general communication pattern, and thus disrupt it, even though it is not possible to know exactly who is communicating with whom. This criticism is also true of xor-trees [10]. As we shall see in the next section, there are several other shortcomings with the bus schemes proposed in [2].

Our Results: The aforementioned problems with deterministic busing lead us to explore further the randomized busing protocol proposed in [2]. In this protocol, messages are delivered by a single bus traveling on a random walk in G. If, for instance, G is complete then the expected delivery time is $\Theta(n)$. We show that if G is complete and there are $w \leq n$ buses, then the expected delivery time for a message can be reduced to $O(n/\sqrt{w})$. We further show that this result is tight—that the expected delivery time is lower bounded by $\Omega(n/\sqrt{w})$. This is somewhat surprising, as one might hope for linear speed up; i.e., a bound of $\Theta(n/w)$. We then define a new class of graphs called r-partite directed collars and we obtain analogous bounds on the delivery time for this class of graphs. We also propose several new features that overcome problems in the original bus system. We show that for an appropriate choice of parameters these new features do not impact the expected delivery time in the case of a complete graph.

2 Background

Before we present our results, we briefly describe the family of protocols presented in [2], which our method builds upon. To get complete details, the reader should see the original paper. The basic idea explored in [2] is explained using the metaphor of a public transportation system. We think of the nodes of the communication system as being 'bus stops' and of there being one or more 'buses' that travel from stop to stop. Each bus has 'seats' $s_{i,j}, 1 \leq i \leq n, 1 \leq j \leq n$, each of which can hold a message.

When the bus arrives at node k, seats $s_{k,j}, 1 \leq j \leq n$, are all modified. If node k wishes to send a message to node j, then the message is encoded (we assume a public key cryptosystem, but this is not the only possibility) and placed in $s_{k,j}$. Otherwise $s_{k,j}$ is filled with random bits. A basic assumption is that it is computationally intractable to tell encrypted messages from random bits. Further, node k checks each seat $s_{i,k}, 1 \leq i \leq n$, for incoming messages. Each message $s_{i,k}$ is decrypted. If the result is garbage, then it is ignored. Otherwise, node k receives the message.

Different schemes are distinguished by the number of buses and the patterns in which they travel. The simplest scheme is to have a single bus that follows a Hamiltonian cycle of G. A more communication intensive scheme involves having $2m$ buses traveling at each time step. A bus traverses each edge in each direction. Messages are relayed from bus to bus until they reach their destination. In order for this to work, each node must maintain a routing table that indicates where a message should go next in order to reach a particular destination (in [2] the routes are always shortest paths). An intermediate protocol involves using the preceding method on some

subset of the edges in G (in fact the first scheme mentioned is just the case where the subgraph is a Hamiltonian cycle).

A basic problem with the schemes we have just described is that the path that a message follows through the network is fixed. If an enemy is able to disrupt messages along the path between two nodes (say by cutting an edge completely or replacing selected seats on a bus with random bits), then it can effectively cut communication between them. Another problem is that the schemes described so far require some sort of global control; i.e., nodes must either know how to route messages to their destination, which requires global knowledge of the network, or in the case of a Hamiltonian cycle this cycle must somehow be established, which again requires global knowledge.

To overcome the first problem, Beimel and Dolev proposed routing a bus randomly. The route the bus follows is a random walk on G. Specifically, at each time step, if the bus is at node u, then we pick a neighbor v of u uniformly and randomly, and send the bus along the edge (u, v). This overcomes the problem of edge failure, since a message will simply not travel through disabled edges. As long as G remains connected, a message will eventually reach its destination (with probability one). Randomized busing also eliminates the need for global routing tables to be stored in each node. However, it introduces a number of new problems:

1. The position of the bus is a random variable. When a node wants to send a message, it has to wait for the bus to arrive first. There is no absolute guarantee on how long this will take.

2. The time a message takes to travel from its source to its destination is also a random variable. Although it is possible to show that this travel time is reasonable with high probability, there is no guarantee that a message will ever reach its destination.

3. Some sort of global control is still required to initialize the system; i.e., the nodes have to agree where and when the bus will start traveling.

4. If the bus 'crashes', meaning it reaches a node and the bus or node becomes disabled before the bus departs, either through accident or malicious behavior, then there is no way for the system to rectify or even recognize this situation.

In this paper, we present a number of modifications to the random walk busing scheme that seek to rectify these problems.

First, however, we make a comment about problem 2. Even in the case that buses travel on deterministic paths, and there is no chance of buses being crashed or corrupted, there is some very small probability of mis-communication. This is because we use random bits to fill the unused seats of the bus. There is a small probability that these random bits will decrypt to some message that seems plausible to the receiver. So it is impossible to remove some small probability of system failure.

3 Our Schemes

We assume we are dealing with a "listening adversary," who can monitor all communication links (either statically or dynamically). As in [2], we assume this adversary is honest-but-curious, meaning it cannot change, delete, or add any messages, or change the state of any node. Also, as in [2], we assume semantic security; i.e., messages are encrypted, say by a public key cryptosystem, so that an eavesdropper cannot effectively distinguish between encryptions of any pair of messages.

The Best Expert Versus the Smartest Algorithm

Peter Chen
Department of Computer Science,
Louisiana State University,
Baton Rouge, Louisiana, 70803, USA

Guoli Ding
Department of Mathematics,
Louisiana State University,
Baton Rouge, Louisiana, 70803, USA

June 27, 2003

Abstract

In this paper, we consider the problem of *online prediction using expert advice.* Under different assumptions, we give tight lower bounds on the gap between the best expert and any online algorithm that solves the problem.

Key words. Online algorithm, online prediction, expert advice.

1 Introduction

The problem of *online prediction using expert advice* is for a predictor to predict, along with n other "experts", a sequence $\sigma = \sigma_1, \sigma_2, ..., \sigma_\ell \in \{0,1\}^\ell$. Here, the only assumptions we make are the following.

Assumption 1.1 *Before predicting each σ_j, the predictor knows the predictions of the experts on this term. Also, right after predicting each σ_j, the predictor is given the true value of this term.*

Notice that we do not assume anything on possible patterns of either σ or the sequences of the predictions of the experts. The goal of the predictor is to "score" as close to the best expert as possible. We point out that this is different from the goal that tries to predict as accurate as possible, which is a problem studied in [1].

Next, we make the problem more precise. Suppose $x = x_1, x_2, ..., x_\ell$ is a sequence of predictions, made by the predictor or by the experts. It is worth mentioning that, sometimes, terms of x might be allowed to take any value in the interval $[0,1]$. The *loss* of x is defined to be

$$L(x,\sigma) = \sum_{j=1}^{\ell} |x_j - \sigma_j|.$$

Let γ_i be the sequence of predictions made by expert i and let $\Gamma = \{\gamma_i : 1 \le i \le n\}$. Then

$$L(\Gamma,\sigma) = \min\{L(\gamma_i,\sigma) : 1 \le i \le n\}$$

is the loss of the best expert. For any strategy $\mathcal{A}$ of the predictor,[1] let $\tau_{\mathcal{A}}(\sigma,\Gamma)$ be the sequence of predictions generated according to $\mathcal{A}$. We measure the performance of $\mathcal{A}$ by the worst *gap* between the losses of the predictor and the best expert. That is, by

$$G_{\mathcal{A}}(n,\ell) = \sup_{\sigma,\Gamma} \left(L(\tau_{\mathcal{A}}(\sigma,\Gamma),\sigma) - L(\Gamma,\sigma)\right).$$

[1]Here we assume that, when predicting two sequences σ' and σ'', if σ' and σ'' turn out to be the same, and the predictions of the experts are also the same, then the strategy should generate two identical sequences of predictions. For a strategy that generates two different sequences of predictions, due to randomization or similar reasons, we will consider it as several strategies (under our term) and this change does not affect our discussions.

Clearly, the goal of the predictor is to minimize $G_{\mathcal{A}}(n,\ell)$ over all strategies $\mathcal{A}$. In this paper, we analyze upper and lower bounds of $G_{\mathcal{A}}(n,\ell)$.

To get a lower bound, let us make the following assumption, which is more in favor of the predictor, and thus will make the result stronger:

Assumption 1.2 *The predictions of the predictor can be any real number in the interval* $[0,1]$, *while the predictions of the experts can only be* 0 *or* 1. *In addition, before predicting* σ_1, *the predictor knows not only* ℓ, *but also* Γ, *the entire prediction sequence of each expert. In other words, other than the actual value of each* σ_j, *the predictor knows everything else before predicting* σ_1.

Under Assumption 1.2, it is proved in [2] that, for all strategies $\mathcal{A}$,

$$\liminf_{n\to\infty} \liminf_{\ell\to\infty} \frac{G_{\mathcal{A}}(n,\ell)}{\sqrt{(\ell/2)\ln n}} \geq 1. \tag{1.1}$$

We should point out that, because of the order the two limits are taken, it is assumed, implicitly, in the above inequality that ℓ is significantly larger than n. We will see later that the situation is quite different if ℓ is smaller than n.

For upper bounds, let us make a different assumption, which is less in favor of the predictor, and thus will make the result stronger.

Assumption 1.3 *The predictions of the predictor and the experts can be any real number in* $[0,1]$. *The predictor also knows* ℓ *before predicting* σ_1.

Under Assumption 1.3, an online algorithm (a strategy for the predictor) $\mathcal{A}$ is given in [2] for which

$$\liminf_{n\to\infty} \liminf_{\ell\to\infty} \frac{G_{\mathcal{A}}(n,\ell)}{\sqrt{(\ell/2)\ln n}} \leq 1. \tag{1.2}$$

In fact, what has been proved is that, for all positive integers n and ℓ, the algorithm $\mathcal{A}$ satisfies

$$G_{\mathcal{A}}(n,\ell) \leq \sqrt{\frac{\ell\ln(n+1)}{2}} + \frac{\log_2(n+1)}{2}. \tag{1.3}$$

Notice that (1.3) is much stronger than (1.2) since it upper bounds $G_{\mathcal{A}}(n,\ell)$ for all n and ℓ. Having such a bound is important because very often, in various applications, n and ℓ are not arbitrarily large. In this paper, we improve lower bound (1.1) in the same way.

Theorem 1.1 *Under Assumption 1.2, for any algorithm* $\mathcal{A}$, *any integer* $n \geq 2$, *and any* $\epsilon \in [0,1]$, *if*

$$\ell \geq \ell(n) := \sqrt{\pi/8}\ ((\ln n)^2 + 8)\ (\sqrt{2\ln n} + 1)\ n^{1-\epsilon}, \tag{1.4}$$

then

$$G_{\mathcal{A}}(n,\ell) \geq \left(\sqrt{\frac{(1-\epsilon)\ell\ln n}{2}} - \frac{1}{2}\right)\ (1 - \delta^n - (1-\delta)^n), \tag{1.5}$$

where

$$\delta = \frac{1}{\sqrt{2\pi}\ (\sqrt{2\ln n} + 1)\ n^{1-\epsilon}} - \frac{(\ln n)^2 + 8}{4\ell}.$$

First, as easily shown below, (1.1) is a consequence of Theorem 1.1, and thus our theorem is indeed an improvement of (1.1) (in the sense that our result implies (1.1) yet it is not implied by (1.1)).

Corollary 1.1 *Inequality (1.1) holds for all online prediction algorithms* $\mathcal{A}$.

Proof. By setting $\delta_1 = 1/(\sqrt{2\pi}(1+\sqrt{2\ln n}\,)n^{1-\epsilon})$, it is straightforward to verify that

$$\begin{aligned}\liminf_{n\to\infty}\ \liminf_{\ell\to\infty}\frac{G_{\mathcal{A}}(n,\ell)}{\sqrt{(\ell/2)\ln n}} &\geq \liminf_{n\to\infty}\ \liminf_{\ell\to\infty}\left(\sqrt{1-\epsilon}-\frac{1}{\sqrt{2\ell\ln n}}\right)(1-\delta^n-(1-\delta)^n)\\ &= \liminf_{n\to\infty}\sqrt{1-\epsilon}(1-\delta_1^n-(1-\delta_1)^n)\\ &= \sqrt{1-\epsilon},\end{aligned}$$

holds for all $\epsilon \in (0,1)$, and so (1.1) follows. ∎

Further remarks on Theorem 1.1.

(a) If $n = 1$, it is easy to see that the algorithm that copies the only expert will perform exactly the same as the best expert and thus $G_{\mathcal{A}}(n,\ell) = 0$, for all ℓ. Because of this, we can say that the assumption $n \geq 2$ in the theorem does not miss any interesting cases.

(b) Inequality (1.5) still holds if we set $\epsilon = 0$. We introduce this extra parameter because we need it in proving Corollary 1.1.

(c) For any $\epsilon > 0$, it is clear that $\ell(n)/n \longrightarrow 0$, as $n \longrightarrow \infty$. Therefore, the requirement $\ell \geq \ell(n)$ is more or less the same as $\ell \geq n$.

(d) One may wonder if the requirement $\ell \geq \ell(n)$ can be dropped completely. For instance, one may ask if there could exist a constant $c > 0$, a function $d(n,\ell)$ with $\lim_{n\to\infty} d(n,\ell) = 0$, and such that

$$G_{\mathcal{A}}(n,\ell) \geq \sqrt{\frac{\ell\ln n}{2}}\,(c + d(n,\ell)) \tag{1.6}$$

holds for all n, ℓ, and $\mathcal{A}$. Unfortunately, the answer is negative. Consider the strategy $\mathcal{A}$ that predicts $1/2$ all the time. Then $L(\tau_{\mathcal{A}}(\sigma,\Gamma),\sigma) = \ell/2$, for all σ and Γ. It follows that

$$G_{\mathcal{A}}(n,\ell) = \sup_{\sigma,\Gamma}\,(L(\tau_{\mathcal{A}}(\sigma,\Gamma),\sigma) - L(\Gamma,\sigma)) = \frac{\ell}{2} - \inf_{\sigma,\Gamma}\,L(\Gamma,\sigma) \leq \frac{\ell}{2}. \tag{1.7}$$

Clearly, this inequality contradicts (1.6), for every $\ell > 0$, when n is sufficiently large. This contradiction indicates that a condition similar to $\ell \geq \ell(n)$ is required to prove any lower bound of the form (1.6).

So far we have discussed the situation when ℓ is bigger than n. When ℓ is smaller than n, we have seen from Remark (d) that lower bounds (1.1) and (1.5) no longer hold. Moreover, as indicated by our next result, that upper bounds (1.2) and (1.3) are not very close to the truth either.

Theorem 1.2 *For all n, ℓ, and $\mathcal{A}$, under Assumption 1.2, we have*

$$G_{\mathcal{A}}(n,\ell) \geq \frac{\ell}{2}(1-(1-2^{-\ell})^n - (2^{-\ell})^n).$$

By combining this result with (1.7) we obviously have the following.

Corollary 1.2 *For all ℓ, under Assumption 1.2,*

$$\lim_{n\to\infty}\ \inf_{\mathcal{A}}\ G_{\mathcal{A}}(n,\ell) = \frac{\ell}{2}.$$

This result suggests that, if n is significantly larger than ℓ, then the predictor cannot catch up with the best expert. The only thing the predictor can do is to predict $1/2$ all the time so that it won't be left too far behind the best expert.

THE WHOLE GREATER THAN THE SUM OF ITS PARTS: COMBINING THE STRENGTHS OF HEURISTIC OPTIMIZATION ALGORITHMS*

Nigel Gwee and Peter P. Chen
Computer Science Department
Louisiana State University
Baton Rouge, LA 70803

Abstract: We describe a general procedure that enables us to combine the strengths of selected heuristic optimization algorithms to produce solutions that are often better than what each algorithm could produce individually. We illustrate our procedure on the Generalized Set Covering Problem. By combining several heuristic algorithms in our procedure, we obtain optimal solutions in many instances. The algorithms used in this way are shown to be effective also in solving the classical Set Covering Problem.

1. INTRODUCTION

Heuristic algorithms are the only practical solution for NP-hard problems. Offering a compromise between solution optimality and speed, these algorithms nevertheless vary in the quality of their solutions from one problem instance to another: where one performs well in some instance, another performs less well, with the reverse holding true for another instance. We describe here a simple general procedure whereby we can combine a set of chosen heuristic algorithms so that we can exploit the best qualities of each algorithm in each problem instance, and produce output that is at least as good as what each algorithm could generate individually. As we shall show, our procedure involves more than just applying each of these algorithms and selecting the best output from among them.

To illustrate our procedure, we solve the *generalized set covering problem* (GSCP) described by Chen and Ding in [CD03b]. The heuristic algorithms we shall implement are: the greedy algorithm (GSCA) proposed by Chen and Ding; a new algorithm that operates in a reverse direction from the greedy algorithm and which we call the "generous algorithm" ; and a more sophisticated version of both these algorithms ("Super Greedy" and "Super Generous").

We show that our procedure produces solutions that are at least as good as any produced by the individual heuristic algorithms. We then apply our algorithms on the classical *set covering problem* and show how the combined forces of our generalized algorithms can be made just as effective as other more specialized algorithms.

1.1. The Generalized Set Covering Problem (GSCP)

The *set covering problem* (SCP) has been extensively analyzed, and numerous efficient algorithms have been presented for its solution. As a representative of the NP-hard problems, SCP has attracted attention because of its application to many real-world problems. Recently, Chen and Ding formulated a generalized version of SCP, which they call the *generalized set covering problem* (GSCP) [CD03b]. This generalized version forms a prototype for applications such as profiling[CD03a].

To facilitate the definition of GSCP and to show its relation to SCP, we first define SCP in a format slightly different from the customary definition. Note that our definition of SCP

* This research is supported by AFOSR Grant No. F49620-01-1-0264.

describes the unweighted form, in contrast with the form adopted by Beasley [BC96], Caprara [CFT98] and others.

Let us first define the symbols we shall be using. Let $\boldsymbol{S}$ be a finite collection of sets, then define $\overline{\boldsymbol{S}}$ to be the union of all members of $\boldsymbol{S}$. Given a function w:$S \rightarrow R_+$, the set of non-negative reals, then for any finite $S' \subseteq S$, define $w(S') = \sum_{s \in S'} w(s)$.

Given a finite set S and $\boldsymbol{S} = \{S_1, S_2, \ldots, S_n\}$, where $S_i \subseteq S$, $1 \leq i \leq n$, we call $\boldsymbol{A} \subseteq \boldsymbol{S}$ a *cover* if $\overline{\boldsymbol{A}} = \overline{\boldsymbol{S}}$.

SET COVERING PROBLEM (SCP); Given a finite set S, and $\boldsymbol{S}$, a collection of subsets of S, find a minimal cover $\boldsymbol{A}$.

We shall refer to the weighted form of the problem (as used by Beasley, Caprara, and others) as the *weighted set covering problem* (WSCP). Here, each covering set is associated with a cost, given by a *weight function* w: $\boldsymbol{S} \rightarrow R_+$. The goal of WSCP is to find a cover $\boldsymbol{A}$ with the minimum total cost ($\boldsymbol{w}(\boldsymbol{A})$). The definition of WSCP follows:

WEIGHTED SET COVERING PROBLEM (WSCP); Given a finite set S, and $\boldsymbol{S}$, a collection of subsets of S, a weight function w: $\boldsymbol{S} \rightarrow R_+$, find a cover $\boldsymbol{A}$ with minimum total cost, $\boldsymbol{w}(\boldsymbol{A})$.

SCP can be redefined as a special case of WSCP, where the weight function is simply w: $\boldsymbol{S} \rightarrow \{k\}$, $k \in R_+ - \{0\}$. This is the basis for calling SCP the *unicost set covering problem*.

The new problem introduced in [CD03b], GSCP, generalizes WSCP in three aspects. First, each $S_i \in \boldsymbol{S}$ is associated with a weighted set $W_i \in \boldsymbol{W}$, where $\boldsymbol{W} = \{W_1, W_2, \ldots, W_n\}$ and $W_i \subseteq W$, $1 \leq i \leq n$, where W is a finite set. Second, each element $s \in S$ is weighted. Third, a combination of weighted elements of S with an additional factor (λ, defined below) enables a relaxation of the covering requirement.

To accommodate the first generalization, we define a weight function c: $W \rightarrow R_+$. Then, by the notation described earlier, for any finite $W' \in W$, $\boldsymbol{c}(W') = \sum_{w \in W'} c(w)$. For any $\boldsymbol{A} \subseteq \boldsymbol{S}$, define the *cost* of $\boldsymbol{A}$, $\boldsymbol{c}(\boldsymbol{A}) = \boldsymbol{c}\left(\bigcup \{W_i : S_i \in \boldsymbol{A}\}\right)$.

To accommodate the second and third generalizations, let d:$S \rightarrow R_+$, and let $\lambda \in [0, 1]$. Then $\boldsymbol{A} \subseteq \boldsymbol{S}$ is called a λ-*d-cover* of $\boldsymbol{S}$ if $d(\overline{\boldsymbol{A}}) \geq \lambda d(\overline{\boldsymbol{S}})$.

GENERALIZED SET COVERING PROBLEM (GSCP): Given S, W, $\boldsymbol{S}$, $\boldsymbol{W}$, d, c, λ, find a λ-d-cover of $\boldsymbol{S}$, $\boldsymbol{A} \subseteq \boldsymbol{S}$, with minimum cost $\boldsymbol{c}(\boldsymbol{A})$.

2. ALGORITHMS FOR GSCP

Chen and Ding have proposed a polynomial-time greedy algorithm (GSCA) for GSCP [CD03b]. We describe the algorithm below, and consider two different cost functions to determine the selection process.

2.1. Greedy Algorithms

Chen and Ding's GSCA modifies Chvátal's algorithm for SCP [Chv79] to accommodate the generalizing parameters.

Algorithm GSCA
Input: $\boldsymbol{S}$, $\boldsymbol{W}$, d, c, λ
Output: $\boldsymbol{A} \subseteq \boldsymbol{S}$, $d(\overline{\boldsymbol{A}}) \geq \lambda d(\overline{\boldsymbol{S}})$

1. Initialize:
    1.1. $\boldsymbol{A} \leftarrow \varnothing$
    1.2. **for** i from 1 to $|\boldsymbol{S}|$ **do**
        1.2.1. $S'_i \leftarrow S_i$
        1.2.2. $W'_i \leftarrow W_i$
2. **while** $d(\overline{\boldsymbol{A}}) < \lambda d(\overline{\boldsymbol{S}})$ **do**
    2.1. $i\text{-}min \leftarrow i$: Cost $(\boldsymbol{S}, \boldsymbol{A}, S_i, W_i) = \min$ [Cost $(\boldsymbol{S}, \boldsymbol{A}, S_j, W_j)$:$S_j \in \boldsymbol{S} - \boldsymbol{A}$]
    2.2. Update:
        2.2.1. $\boldsymbol{A} \leftarrow \boldsymbol{A} \cup \{S_{i\text{-}min}\}$
        2.2.2. **for** each $S_k \in \boldsymbol{S} - \boldsymbol{A}$ **do**
            2.2.2.1. $S'_k \leftarrow S'_k - S_{i\text{-}min}$
            2.2.2.2. $W'_k \leftarrow W'_k - W_{i\text{-}min}$

Algorithm Cost_1
Input: $\boldsymbol{S}$, $\boldsymbol{A}$, $S_j \subseteq S$, $W_j \subseteq W$
Output: *cost*

1. **if** $d(S_j) = 0$ **then**
    $cost \leftarrow \infty$
  else if $d(\overline{\boldsymbol{A}} \cup S_j) \leq \lambda d(\overline{\boldsymbol{S}})$ **then**
    $cost \leftarrow \boldsymbol{c}(W_j) / d(S_j)$
  else
    $cost \leftarrow \boldsymbol{c}(W_j) / (\lambda d(\overline{\boldsymbol{S}}) - d(\overline{\boldsymbol{A}}))$

A simple alternative to the above Cost function is to take $\boldsymbol{c}(W_j)$ alone as the cost:

Algorithm Cost_2
Input: $\boldsymbol{S}$, $\boldsymbol{A}$, $S_j \subseteq S$, $W_j \subseteq W$
Output: *cost*

1. **if** $d(S_j) = 0$ **then**
    $cost \leftarrow \infty$
  else
    $cost \leftarrow \boldsymbol{c}(W_j)$

This simpler computation is often sufficient. In our simulations, we shall compare the performance of these two Cost functions.

This algorithm sometimes results in solutions that contain redundant elements in the cover, i.e., the cover remains a cover after the removal of these elements. An enhancement of the algorithm will be to add a post-processing phase wherein we reduce the solution as much as we

4Conclusion

The CIPIA Fellow program at LSU was very successful. Dr. Steve Seiden (a CIPIA Fellow) was very promising and very productive. However, he was involved in an accident and passed away at a very young age. Otherwise, we would have seen him become a very productive researcher in CIPIA. Teaming up with his mentor (Dr. Peter Chen), Dr. Guoli Ding (a CIPIA Fellow) has successfully obtained a large 5-year NSF grant ($1.8 million dollars) to work in the area of cyber security as a co-PI. He also got promoted to Full Professor at the Math Department at LSU. Even though Dr. Gwee started late in the Fellow program, he is now a part-time post-doctor research in an NSF research project in CIPIA and has been incorporating CIPIA material into his courses attended by undergraduate and graduate students.